DOCUMENT RESUME

ED 139 854 UD 016 958

AUTHOR Anderson, Richard B.
TITLE The Effectiveness of Follow Through: What Have We Learned?
INSTITUTION Abt Associates, Inc. Cambridge, Mass.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE 5 Apr 77
CONTRACT 300-75-0134
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, N.Y., April 5, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Gains; Achievement Tests; *Compensatory Education Programs; Elementary Education; *Failure Factors; *Minority Group Children; Models; Program Effectiveness; *Program Evaluation; *Success Factors; *Test Results
IDENTIFIERS *Project Follow Through

ABSTRACT

This paper presents the results of a five year effort to evaluate the results of tests and questionnaires administered to Follow Through (FT) and non-Follow Through (NFT) children and their parents and teachers. The test battery studied included the Metropolitan Achievement Test (MAT), the Raven's Progressive Matrices instrument, the Cooperville Self Esteem Inventory, and the Intellectual Achievement Responsibility Scale. For each of 139 distinct FT groups, distributed unevenly among 17 models, performance on various post tests was compared with national norms and with statistically constructed expected performance levels for similarly disadvantaged groups outside the program. The results of these comparisons indicated that 27 or 19% of the groups performed up to national norm standards on MAT. Overall, the positive effects of FT on scores of the MAT seemed to outnumber its negative effects for only a minority of the groups that received FT service. In general, across all models, all groups, and all measures, there were fewer positive effects (12.8%) than negative (19.6%) and a preponderance of null effects (67.6%). Of all the findings, the most pervasive, consistent, and suggestive was this: the effectiveness of each FT model depended more on local circumstances than on the nature of the model. (Author/AM)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED139854

ABT ASSOCIATES INC.
55 WHEELER STREET, CAMBRIDGE, MASSACHUSETTS 02138
TELEPHONE • AREA 617-492-7100
TELEX: 710-320-6367

THE EFFECTIVENESS OF FOLLOW THROUGH: WHAT HAVE WE LEARNED?

by

Richard B. Anderson
Abt Associates Inc.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

UD016958

Presented at
the Annual Meeting of the
American Educational Research Association
at New York
5 April 1977

This paper is based on research performed by Abt Associates Inc. under Contract No. 300-75-0134 to the United States Office of Education. This Research is the product of many hands; neither Abt Associates staff nor the Office of Education is responsible for any errors or omissions: the author accepts full responsibility for those. The conclusions in this paper do not necessarily reflect the final judgments of the Office of Education; these will appear in USOE's forthcoming report to Congress on the National Follow Through Evaluation.

THE EFFECTIVENESS OF FOLLOW THROUGH: WHAT HAVE WE LEARNED?

Over the past decade, more than twenty teams of educational experts, representing a wide variety of theoretical and practical opinions and approaches, have served as "sponsors" of "models" of compensatory intervention in the early primary grades. Together, they comprise the national Follow Through experiment. As part of the national evaluation of Follow Through, Abt Associates has now spent nearly five years studying the results of tests and questionnaires administered to Follow Through (FT) and non-Follow Through (NFT) children and their parents and teachers, in the hope of discovering how effective the various approaches have been.

As we complete the process, we are aware that our results* provide a unique perspective on the performance of Follow Through, both as a program of compensatory service and as an experiment which is intended to teach us how to run better programs and better research in the future. We are also aware that our results do not constitute a comprehensive evaluation of the national Follow Through endeavor: they partake, rather, of the nature of a report card.

This analogy--albeit flawed, as every analogy must ultimately be--is neither frivolous nor trivial. When my young son brings home a report care that indicates that he is reading better than he did last term and getting along better with his peers, I am pleased and make a point of letting his know it in some very concrete ways. When it also says that he seems preoccupied and cannot keep his mind on his arithmetic work, I look for an explanation to my fuller knowledge of the child. I may or may not bother to tell the teacher that the lad has been working out an acute sibling rivalry with his older brother, or that he has reacted very strongly to the death of a friend or the loss of a favorite toy. For my own parental purposes, I take the report card as one source of information among many which help me to make informed choices as I try to help him over the hurdles that lie between him and a fulfilling adulthood.

---

*Abt Associates Inc., Education as Experimentation: A Planned Variation Model, Volume IV.

Similarly, we are anxious that nobody take our findings out of context. Much of the existing evidence on the worth and consequences of Follow Through intervention lies outside our data base and therefore outside our "report card." Other evaluation reports, including the self-evaluations of the sponsors, can come closer than we to providing a complete catalogue of all that has been good and bad, effective and ineffective about Follow Through in its decade of development and trial. Those who have practical decisions to make about Follow Through as a program would be mistaken to base their decisions on only a part of the available evidence, whether our part or another. They would be equally mistaken to ignore our results: seen in the proper context, and interpreted judiciously, they embody a unique part of the history from which we must learn if we are to build upon our collective experience.

The report card analogy breaks down, of course, in some very important particulars. I do not have to decide whether to continue supporting and working with my son: his intrinsic worth and importance transcend mere evidence. Many who have a stake in Follow Through evince a similar absolutism of feeling toward the program or toward aspects of it in which they have invested particularly heavily. The nation, however, through the agency of the Congress, has invested more than half a billion dollars of tax revenues in Follow Through in the expectation that certain benefits would accrue to disadvantaged children. Where the benefits are evident, responsible policy-makers must judge whether they were worth the cost, and whether further investment is called for. Should the expected benefits not appear, the nation could legitimately decide to spend its resources elsewhere.

The data with which we have worked were gathered in order to see how well Follow Through would deliver a variety of benefits measurable by more or less traditional testing procedures. The Metropolitan Achievement Test provided measures of such academic skills as reading, arithmetic, spelling, vocabulary, and language. The Raven's Progressive Matrices instrument tested abstract problem-solving ability. The Coopersmith Self-Esteem Inventory and the Intellectual Achievement Responsibility Scale addressed the more affective domains of self-concept and the tendency to attribute success and failure to oneself or to others. While by no means covering the full range of benefits Follow Through aimed to provide, this set of

measures is not without breadth. Many parents, educators, taxpayers, and legislators tend, furthermore, to agree that the measured domains are important to children's academic success. The test battery does not measure perfectly what it measures, but it probably comes as close as could anything available ten years ago--or now, for that matter.

As measured by these tests, then, what benefits might one have expected Follow Through intervention to deliver? Depending on one's beliefs about disadvantage and about intervention, one might wish to impose any of a number of criteria at three general levels of stringency:

- Achievement equality: to succeed, Follow Through must eliminate the academic consequences of socioeconomic disadvantage, raising poor children to the performance levels of their more advantaged peers.

- Positive treatment effect: to succeed, Follow Through must help children to perform better than they would have done without the program.

- Non-negative treatment effect: to succeed, Follow Through must at least leave children's test performance levels where they would have been without treatment, while presumably providing untested benefits of long-term importance.

In the heady atmosphere of the War on Poverty, the program grew out of the perception that it was not fair for children to suffer academically merely because they were poorer than other children. Many educators had ideas about how to help children do better in school, and, presumably, in life, but most of these ideas had never had the resources for a fair trial. It seemed plausible that at least some of these ideas, realized in Follow Through models, might in practice yield really impressive results, perhaps even up to the level of achievement equalization. Since there was no way to know a priori which models would work, Follow Through Planned Variation provided a vehicle for a wide variety of them. The Head Start experience had already demonstrated, however, that the educational consequences of socioeconomic disadvantage are diverse and tenacious. In view of that experience, one might be content with demonstrable progress toward achievement equality: that is, positive treatment effects. Given, moreover, the well-documented tendency of poor children to fall further and further behind as they grow

*A "group" in these discussions comprises all the FT children in a model's local project in one of the four cohort streams (Cohorts II and III, each with entering kindergarten (K) and entering first (EF) streams). Responsive Education's Cohort III-K children in Duluth, for example, constitute a "group." The groups are the units of the effect averages reported later in this paper.

older, one could even argue that a program might be successful if it simply kept them from losing ground.

Within the limitations of the test battery, the data contain the information necessary to judge the performance of Follow Through at each of the three levels. For each of 139 distinct Follow Through (FT) groups,* distributed rather unevenly among 17 models, we compared performance on the various posttests with national norms and with statistically-constructed expected performance levels for similarly-disadvantaged groups outside the program. The results of these comparisons provide answers to several evaluative questions; let us begin with the most demanding ones:

Question 1: Did Follow Through raise poor children's test scores to national norm levels?

Figure 1 provides the first half of our answer to this question; it shows that 27 (19%) of the groups performed up to national norm standards on one or more of the three subtests shown. Whatever Follow Through did for the children it served, by and large it did not make them indistinguishable from the national average: at posttest, the Follow Through population is still identifiably disadvantaged in the sense of below-average scores on nationally-standardized tests of traditional academic achievement.

But what of the 27 groups that did score as well as the rest of the nation on one or more tests? Was it Follow Through that pushed them to that good performance, or were these merely the least disadvantaged of the Follow Through groups, so that they would have done well even without the treatment? Figure 2 addresses this question by imposing a second requirement on the groups: here, the dots denote groups that scored above national norms and for whom we have evidence that they scored better than one would expect similarly-disadvantaged groups to score without Follow Through help. Whereas 11 models had at least one group that attained national norm performance on one or more tests, only five had groups that met both criteria, and only one of those had more than one such group. The most successful model, in this respect, achieved this heroic level of performance with only a minority of its groups. If educators know how to close the achievement gap between disadvantaged children and the nation in general, we must conclude on this evidence that Follow Through has not proved that they can do it reliably.

Figure 1. Frequency with which 139 Follow Through groups scored at or above national norms* on three MAT posttests, by Follow Through model.

MODEL	NORMED MAT SUBTESTS			
	Reading	Mathematics	Spelling	Any Subtest
Responsive Education	●●○○○○○○○○○○○○○○	○○○○○○○○○○○○○○○○	●●○○○○○○○○○○○○○○	●●○○○○○○○○○○○○○○
TEEM	●○○○○○○○○○○○○○○○	○○○○○○○○○○○○○○○○	●○○○○○○○○○○○○○○○	●●○○○○○○○○○○○○○
Bank Street	●●●●○○○○○○○○○	●●○○○○○○○○○○○	●○○○○○○○○○○○	●●●○○○○○○○○○
Direct Instruction	●●●●●●●○○○○○○○○○	●●●●●●●●○○○○○○○○	●●●●●●○○○○○○○○○○	●●●●●●●●○○○○○○○○
Behavior Analysis	●●○○○○○○○○○○○○	●●○○○○○○○○○○○○	●●●○○○○○○○○○○○	●●●○○○○○○○○○○○
Cognitive Curriculum	●○○○○○○○○	○○○○○○○○○	●○○○○○○○○	●●○○○○○○○
Parent Education	●●○○○○○○○○○○○	○○○○○○○○○○○○○	●●○○○○○○○○○○○	●●●○○○○○○○○○○
EDC	●○○○○○○○○○	●○○○○○○○○○	●○○○○○○○○○	●○○○○○○○○○
Self-Sponsored	○○○○○	○○○○○	●○○○○	●○○○○
IELP	●○○	●○○	●●○	●●○
Interdependent Learning	○○○	○○○	○○○	○○○
SEDL	○○○○○○○○	○○○○○○○○	○○○○○○○○	○○○○○○○○
Cultural Linguistic	○○	○○	○○	○○
Home-School Partnership	○○	○○	○○	○○
California Process	○○○○○○○○	○○○○○○○○	○○○○○○○○	○○○○○○○○
New School Approach	●○	○○	●○	●○
Mathemagenic Activities	○○○	○○○	○○○	○○○

KEY: ● At or above national norm level
○ Below national norm level

*A group was judged to have scored at or above national norms if the proportion of its members scoring one or more years behind grade level was no greater than the corresponding national proportion.

Figure 2. Frequency with which 139 Follow Through groups scored at or above national norms* and also better than expected of similarly-disadvantaged groups outside Follow Through, by model.

MODEL	NORMED MAT SUBTESTS			Any Subtest
	Reading	Mathematics	Spelling	
Responsive Education	oooooooooooooo	oooooooooooooo	oooooooooooooo	oooooooooooooo
TEEM	oooooooooooooo	oooooooooooooo	oooooooooooooo	oooooooooooooo
Bank Street	•ooooooooooo	•ooooooooooo	ooooooooooo	•ooooooooooo
Direct Instruction	••••oooooooooooo	•••••ooooooooooo	•oooooooooooooo	•••••ooooooooooo
Behavior Analysis	ooooooooooooo	ooooooooooooo	ooooooooooooo	ooooooooooooo
Cognitive Curriculum	•oooooooo	oooooooo	oooooooo	•oooooooo
Parent Education	ooooooooooo	ooooooooooo	•ooooooooooo	•ooooooooooo
EDC	oooooooooo	oooooooooo	oooooooooo	oooooooooo
Self-Sponsored	ooooo	ooooo	ooooo	ooooo
IELP	•oo	•oo	•oo	•oo
Interdependent Learning	ooo	ooo	ooo	ooo
SEDL	oooooooo	oooooooo	oooooooo	oooooooo
Cultural Linguistic	oo	oo	oo	oo
Home-School Partnership	oo	oo	oo	oo
California Process	oooooooo	oooooooo	oooooooo	oooooooo
New School Approach	oo	oo	oo	oo
Mathemagenic Activities	ooo	ooo	ooo	ooo

KEY: • At or above national norm level, and also above expectation
o Below national norm level, or no higher than expectation

* A group was judged to have scored at or above national norms if the proportion of its members scoring one or more years behind grade level was no greater than the corresponding national proportion. It was judged to have scored better than expectation if its adjusted Follow Through/non-Follow Through comparison indicated a significant and material positive effect of program participation (cf. p. 17).

Question 2: Did Follow Through raise poor children's test scores?

If Follow Through didn't close the disadvantage gap entirely, did it at least narrow it? Did Follow Through children score better than they would have done if the program had not intervened? To address this (strictly non-empirical) question, we are forced into a variety of statistical approaches to estimating the test performance of "similarly-disadvantaged" children outside the Follow Through program. Where the data permit, we have compared the scores of each Follow Through group with an estimate based on those of its local non-Follow Through (NFT) comparison group and also with another estimate based on those of all the NFT children tested. When we speak of an "effect," we are referring to the result of one such comparison.

In this context, it is also useful to turn from the traditional, subject-matter-oriented groupings of outcomes to groupings that reflect more of the differences among the emphases of the models. We have divided the eleven tests and subtests into three broad domains, as shown in Figure 3. An indication of the distribution of "positive effects" appears in Figure 4. Here, each dot denotes a Follow Through group that scored higher more often than lower, on the tests of the indicated domain, in comparison with various estimates of non-Follow Through performance. Overall, the positive effects of Follow Through on these test scores seem to outnumber its negative effects for only a minority of the groups that received Follow Through services. No model shows a net positive balance in all its groups, and only one in even a clear majority of them, in every domain.

Question 3: Did Follow Through at least leave children's test scores as high as they would have been without intervention?

In general, across all models, all groups, and all measures, we find fewer positive effects (12.8%) than negative (19.6%) and a preponderance (67.6%) of null effects.* On the evidence, it is not at all uncommon for a group of Follow Through children to score less well on a test than one would expect them to score if they had not received the treatment.

---

*Where differences were too small to seem educationally important, or where they were statistically insignificant and thus could easily be explained by differences among children within groups, we declared them null. Our

FIGURE 3

DOMAINS OF THIRD GRADE TESTING IN FOLLOW THROUGH

Outcome Domains	Tests	
BASIC SKILLS	Word Knowledge Spelling Language Math Computations	Metropolitan Achievement Test (MAT)
COGNITIVE CONCEPTUAL SKILLS	Reading Math Concepts Math Problems	Metropolitan Achievement Test (MAT)
	Raven's Progressive Matrices	
AFFECTIVE OUTCOMES	Coopersmith Self-Esteem	
	Achievement Responsibility, Positive Achievement Responsibility, Negative	Intellectual Achievement Responsibility Scale (IARS)

---

criterion of "educational importance" or "material significance" was set at one-fourth of a standard deviation of the outcome variable. For those outcomes that possess national grade-equivalent norms, in the ranges of scores actually achieved by Follow Through children, this criterion corresponds to a difference on the order of two months of achievement. We also required "statistical significance" at the $p < .05$ level, in order to ensure that we did not draw undue conclusions from groups of inadequate size. In general, and most particularly in the entering kindergarten streams, the statistical criterion turned out to be the more stringent of the two; fully 25% of the null effects were "material" but not "significant," whereas only 5% were "significant" but not "material." We surmise that an extra year's attrition in the kindergarten streams may have diminished the numbers of children in various groups sufficiently to reduce the power of our statistical tests in some cases and to create some tendency toward Type II error.

Figure 4. Frequency with which positive effects exceeded negative in three domains, by model.

MODEL	OUTCOME DOMAINS		
	BASIC SKILLS	COGNITIVE CONCEPTUAL	AFFECTIVE
Responsive Education	●●●○○○○○○○○○○	●●●●○○○○○○○○	●●●○○○○○○○○○○
TEEM	●●●○○○○○○○	●●○○○○○○○○○	●●○○○○○○○○○○
Bank Street	●●○○○○○○○○	●●○○○○○○○	●●●○○○○○○○
Direct Instruction	●●●●●●●○○○○	●●●●●●●○○○○○○	●●●●●●●●●●●○○○○
Behavior Analysis	●●●●●●○○○○○	●●●●●○○○○○○○	●●●●●○○○○○○○○
Cognitive Curriculum	●●○○○○○	●●○○○○○	●○○○○○○
Parent Education	●●●●●○○○○○○	●●●●●○○○○○○	●●●●●○○○○○○
EDC	○○○○○○○○○	○○○○○○○○	●●○○○○○○○
Self-Sponsored	●●○○○	●●○○○	●○○○○
IELP	○○○	●○○	●○○
Interdependent Learning	●○○	○○○	○○○
SEDL	●●○○○○○	●●○○○○○○	●●●●●○○○
Cultural Linguistic	●○	○○	○○
Home-School Partnership	○○	○○	○○
California Process	●●●○	●●○○○○	●●●○○○○
New School Approach	○○	○○	○○
Mathemagenic Activities	●○○	●●○	○○○

KEY:
- ○ More positive than negative effects
- ● At least as many negative as positive effects

NOTE: Groups have been excluded from this figure if more than half their effects estimates were judged untrustworthy for methodological reasons.

The data have provided fairly convincing if puzzling answers to our three main questions about the performance of Follow Through, but these answers merely raise another series of questions:

Why are there so few positive effects? Apparently, it is genuinely hard to raise the test scores of disadvantaged children, even in comparison to those of other disadvantaged children, by means of the kinds of compensatory interventions that Follow Through has tried. Why? Several possible answers, more or less consistent with each other, suggest themselves.

1. Disadvantaged children are different from the general population. Much of the philosophy of some models is based on the proposition that different children should have different educational experiences directed at different objectives which imply in turn different criteria for evaluation. Disadvantaged children, some argue, are not defective children but merely different children who should not be judged by the same standards as other children. In our view, this argument amounts to a concession that compensatory education might not have the power to enable its clients to compete successfully with their peers in the nation at large. Be that as it may, our evaluation has undertaken to apply a uniform standard to all the diverse manifestations of Follow Through.

2. It is difficult to implement change in the functioning of institutions by means of intervention from the outside. Efforts to innovate in schools, even spearheaded by well-placed insiders, are often frustrated by the overt or covert unwillingness of teachers, parents, pupils, administrators and others to change accustomed ways of carrying out the business of education. Abundant anecdotal evidence attests to the wide variety of problems that the sponsors have faced in trying to implement their models in the sites. Any extensions of Follow Through to other districts, however, would probably encounter analogous circumstances, and so we are confident that our findings reflect the outcomes one might expect from a broader Follow Through program, even though the available evidence on the particular implementation difficulties to date was too unsystematic to contribute directly to the findings.

3. In some Follow Through models, children get very little practice in test-taking. Several models make it a point to shield children from evaluative testing. Other models, however, provide ample opportunities for

children to develop test-taking skills and develop familiarity with items similar to those found on standardized tests. Therefore, one would not be surprised to find many FT children doing less well on the tests in the outcome battery than do children more accustomed to test-taking.

4. Disadvantage is resilient. Implicit in much of the early thinking which led to Head Start and Follow Through was the idea that disadvantaged children were like a compressed spring: if an educator could merely remove whatever obstacles prevented children from realizing their potential, then they would in the normal course of events expand and stay expanded, barring further negative intervention from outside. When it was alleged that Head Start gains tended to decrease over time, two possible explanations arose: (a) public schooling destroys the benefits of Head Start, re-compressing the spring, as it were; or (b) disadvantaged children are not like compressed springs, but rather like shorter springs that are already in their equilibrium state: if intervention expands them, they tend to return to their normal position in the absence of continuing effort to keep them stretched. Follow Through was designed to cover both of these possibilities, by improving the performance of public schools and by extending the compensatory effort in time. The results of the present Follow Through evaluation effort are consistent with the idea that whatever the forces may be which cause and maintain low test performance among children from families of low socioeconomic status, the compensatory interventions that have been tried to date are not strong enough to countervail them. That is, the models appear to be less effective in raising test scores than originally expected.

Why are there so many negative effects? The advocates of Follow Through did not enter upon the program with the expectation that compensatory services with firm theoretical foundations, implemented by concerned professionals, over a considerable period of time, at substantial costs, and with the active cooperation of the parents of the children being served, could be detrimental to children. How, then, do we explain the fact that we find half again as many negative effects as positive? The question is worth asking, if only to provide another opportunity to make clear what the results really say.

The outcome tests were designed to assess performance with respect to the traditional objectives of schooling, pursued in the usual ways. They are not necessarily the best possible tests of performance with respect to the more innovative objectives of some Follow Through models. In many cases, the measured objectives do not match the objectives of these models. For example, there are models which place little emphasis on reading and math skills, choosing instead to focus on development of a positive sense-of-self. Children associated with these types of models spend their time on activities which are focused on objectives other than development of reading and math skills per se. One's interpretation of our findings will therefore depend on one's educational values and objectives: if one believes that the primary function of schooling is to enhance the ability of children to perform well on the standardized tests in this outcome battery, then the results justify a conclusion that Follow Through has been, about one time in five, a less productive experience than traditional schooling would have been. If, on the other hand, one believes that other broader objectives are more appropriate, and that our instruments do not measure those objectives, then one could speculate--or adduce independent evidence--that these negative results may be offset by progress on other unmeasured outcomes. Our results cannot confirm or deny that possibility; they bear only on the evaluation of Follow Through with respect to traditional criteria.

Why are there so many null effects? To balance the picture, let us also note that most of the effects are null. That is, as determined by the analyses, most Follow Through groups have performed neither better nor worse than expected of similarly disadvantaged non-Follow Through groups on the tests administered at the end of third grade. There are at least three possible explanations for the null effects:

a. FT and NFT may not be very different in practice:

- NFT children generally receive assistance from Title I and other programs, and so we have clearly not compared Follow Through to "the absence of compensatory education."

- In some locations, there is evidence that NFT classrooms have picked up FT techniques and materials from enthusiastic FT participants, thus muddying the practical distinction between treatment and control.

- Finally, implementation difficulties may have kept the Follow Through models from making as much difference as they would have wished.

To the extent that any of these arguments explain the preponderance of null effects, however, the evaluative implications for Follow Through seem rather negative. If Follow Through is worth what it costs, it should be more productive than other, generally less expensive, compensatory programs. If technology transfer in the teachers' lounge is all it takes to wipe out the difference between FT and NFT, then simpler and less expensive dissemination procedures than sponsorship ought to suffice to spread the benefits of Follow Through throughout the nation. And if a program cannot be put into practice, what real value does it have?

b. FT and NFT may be very different, but both ineffective:

Our best evidence of the ineffectiveness of NFT is the oft-cited tendency of disadvantaged children to fall farther and farther behind their more advantaged peers. If this pattern holds for the NFT groups--and we have made no direct test of the hypothesis--then our comparative results suggest that it also holds, and slightly more often, for the FT groups. Our best interpretation of the large number of nulls is the same as for the small number of positive effects: it seems to be hard to make a difference.

c. The tests of differences may not have been powerful enough:

By reason of small sample sizes within some sites, in fact, our analyses were sometimes not sensitive enough to detect all of the effects which may actually exist.

All in all, then, it appears by several standards of judgment that Follow Through has had highly inconsistent and less positive effects on the achievement test scores of children than many of its advocates would have hoped and expected. Merely to intervene is clearly not to guarantee that test scores will rise. This very inconsistency suggests yet another question:

Question 4: Did the various educational strategies tested in Follow Through have different impacts?

Nobody today seriously expects disadvantage, in all its variety, to yield tamely to any known panacea or set of panaceas. One important motiva-

tion for the planned variation approach was the hope that by trying many kinds of intervention in many settings we might be able to learn what works under what circumstances. Do some models or broad types of models have importantly better effects than others? Figure 5 displays, and Chapter Five of Volume IV-A of our report to USOE (cited above) discusses extensively the evidence for several findings that bear upon an answer to this question:

- Models that emphasize basic skills succeeded better than other models in helping children gain these skills.
- Where models have put their primary emphasis elsewhere than on the basic skills, the children they served have tended to score lower on tests of these skills than they would have done without Follow Through.
- No type of model was notably more successful than the others in raising scores on cognitive conceptual skills.
- Models that emphasize basic skills produced better results on tests of self-concept than did other models.

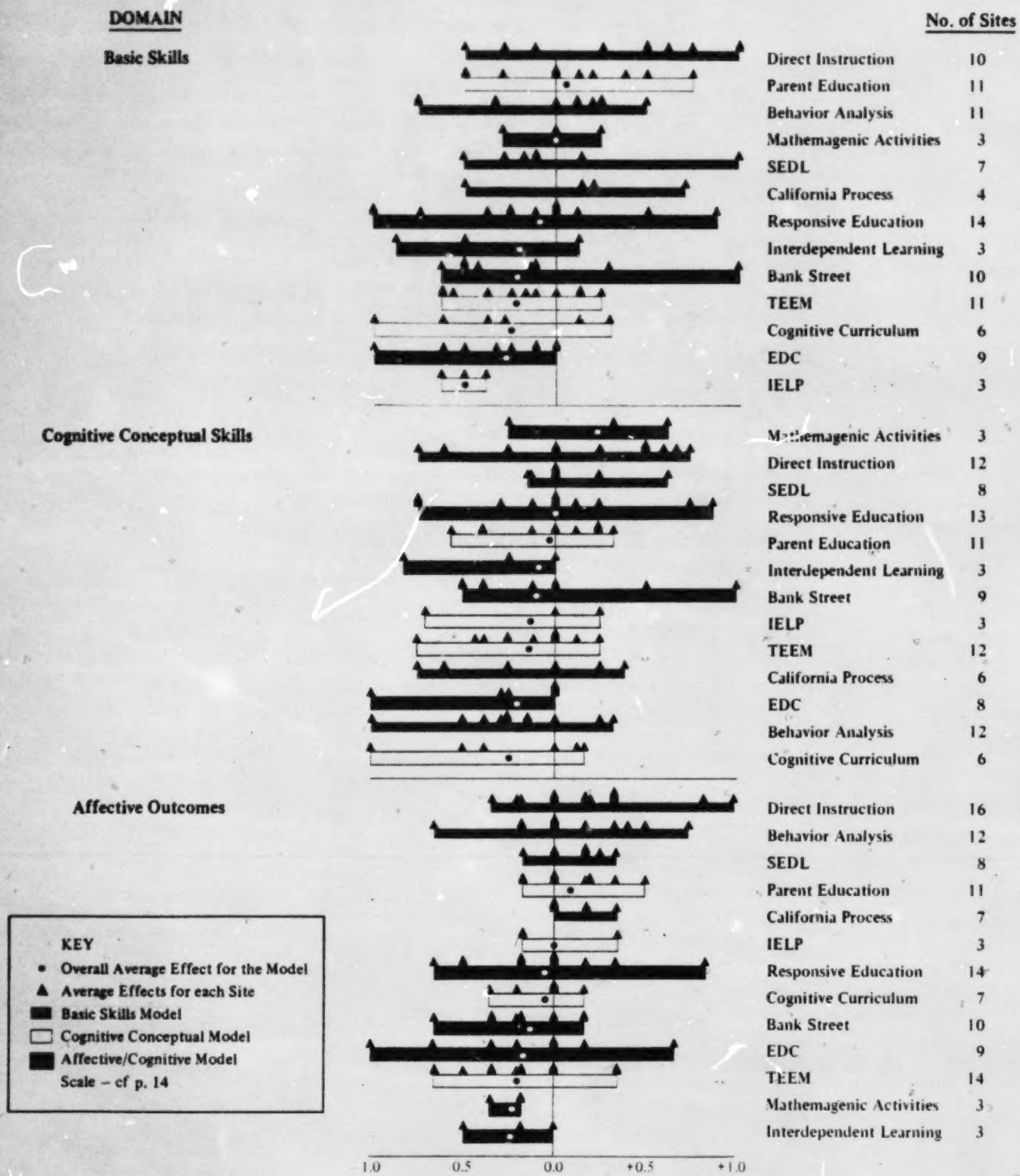
These relative findings seem to reflect very stable patterns in the data. In the context of the earlier findings of general relative ineffectiveness in raising test scores, though, they lose a great deal of their practical value. Of all our findings, the most pervasive, consistent, and suggestive is probably this:

<u>The effectiveness of each Follow Through model depended more on local circumstances than on the nature of the model</u>.

No model has shown itself to be powerful enough to succeed everywhere it has been tried. Figure 5 indicates that eleven models had at least one group with a positive average effect* on basic skills. That is, nearly all models appear to provide a net benefit to some groups of disadvantaged

*Average effects are computed by assigning a value of +1 to each positive effect estimate, -1 to each negative, and zero to each null, and then computing the simple mean of these values. The average effect is thus the number of positive effects, less the number of negative effects, divided by the total number of effects; in other words, it is the proportion of net positive effects. Average effects in the basic skills and cognitive conceptual skills domains are computed on a maximum of eight effect estimates per site/cohort stream group: two estimates of the effect on each of four test scores in each group. In the affective domain, we have only three scores, and so the averages are computed on a maximum of six estimates per group.

FIGURE 5
SITE AND MODEL AVERAGE EFFECTS IN THREE OUTCOME DOMAINS FOR THIRTEEN FOLLOW THROUGH MODELS ACROSS ALL COHORT STREAMS



children relative to other similarly disadvantaged groups. The same finding obtains in the other domains. Every model, on the other hand, had groups with negative average effects. In each model with more than three groups, the range of group average effects was wider than the total range of the thirteen model averages. Each model might have performed very differently in other sites: if each model had been judged by its results in one site only, and that site had been diabolically chosen, the rank order of model averages could have been almost reversed. For this reason, we have taken care not to overinterpret small differences in averages or rankings among models. It is well that the Follow Through planned variation design called for each model to be implemented in a wide variety of places; without this diversity, we would not have been able to demonstrate so clearly the importance of site conditions.

What follows from these results? From time to time we are asked questions which bear upon the believability and generalizability of the patterns we have presented here. Perhaps we can best put our findings into context and delimit the range of their applicability by discussing here some of these questions and our best answers to them:

1. Did we study the right outcomes? Our results comprise evidence on the effectiveness of Follow Through intervention of various kinds in helping disadvantaged children to score better on tests of various kinds in helping disadvantaged children to score better on tests of several academic skills, self-concept, and achievement responsibility than they would have done without Follow Through. Some of the models' sponsors place high priority on the skills tested, and some do not. To the extent that the tests overlap the models' objectives, our results provide evidence on the models' success in attaining those objectives. But whether or not we have anything to say about the models' achievement of their own goals, we do have some important findings about their performance with respect to some rather widely-accepted (if controversial) criteria. We have simply laid several uniform measuring sticks up against each member of a very diverse set of educational situations. Our report describes such apparent program effects as we have discerned in each measured domain; the reader must decide whether there was any reason to expect effects and what judgments about the programs properly follow.

We should also point out that, as in all social sciences--or in any other kind of science, for that matter--the designers of Follow Through had to live with a pervasive trade-off between relevance and measurability. The characteristics one would most like to know can always be measured only indirectly, approximately, and at great cost. Anything the evaluator can actually get at is invariably the "wrong measure," by informed consensus. In our judgment, the Follow Through design and instrumentation represent as good a resolution of this trade-off as the political and methodological climate of the 1960's would give one any reason to expect. Even with the experience of the intervening decade, it is not obvious that anybody is in a position to do importantly better.

2. Were the outcomes measured adequately? Again, the measures in the Follow Through battery were the best available at the time that instrumentation decisions had to be made. Some mid-stream instrument changes were made, at some cost to the longitudinal integrity of the evaluation. As in any large-scale data collection effort, especially a long-term one in a mobile population beset here and there with serious social unrest, the testing and interviewing process encountered many practical difficulties. Abt Associates was not a party to the data gathering, but the quality of the data we have received over the years from the Stanford Research Institute bespeaks a high level of care and ingenuity devoted to overcoming those difficulties.

Here and there throughout the report, we have remarked that the basic skills test scores have tended to correlate more highly with other measured characteristics than have the scores on cognitive conceptual skills, and both more highly than the affective scores. It is not surprising that the simplest and most concrete outcomes, such as spelling, computation, and vocabulary, would be more reliably measured than would more abstract, reactive outcomes such as self-esteem or achievement responsibility. If a child completes several arithmetic computations correctly, he has the skill being measured. If he says he feels good about himself, the response may have any of a number of meanings. Because of such considerations, the prudent reader will place more confidence in the reported effects on the simpler outcomes, but even the patterns in the more complex outcomes are anything but random.

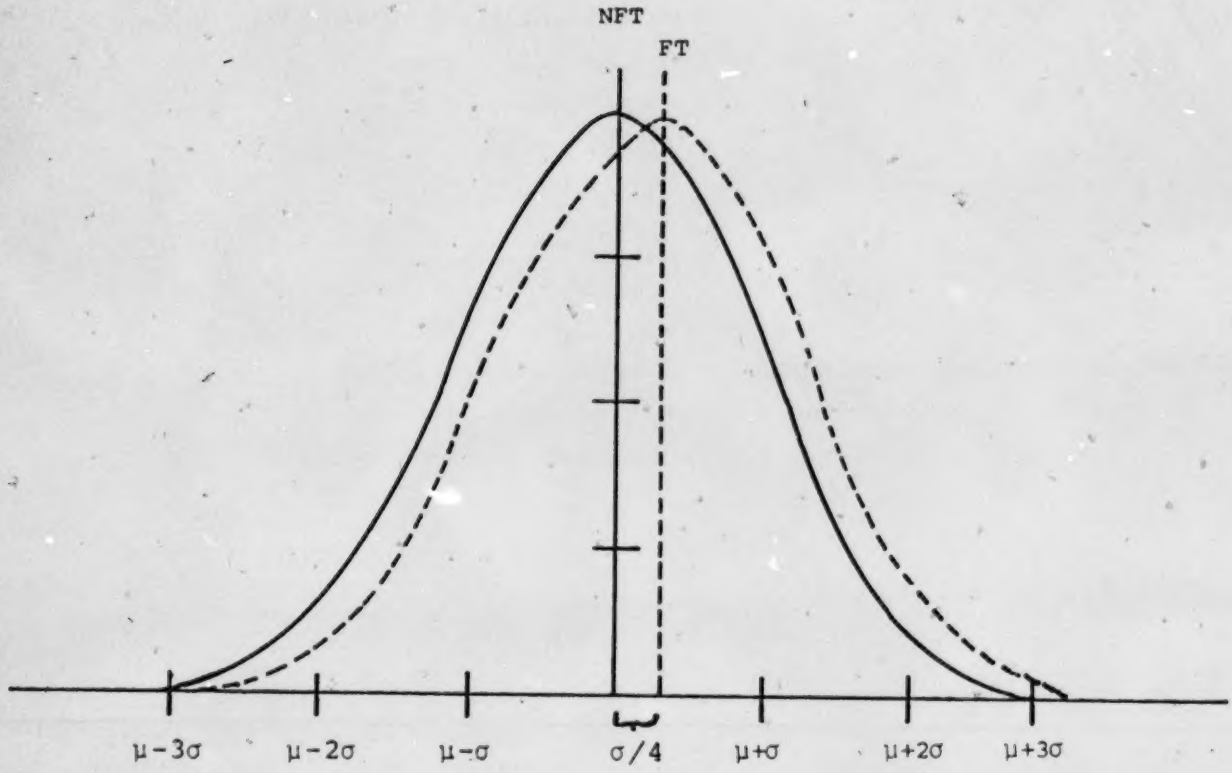
3. Did our analyses set too high a standard of success? Before we would declare an apparent effect "real," whether positive or negative, we required it to be statistically significant (at the conventional $p < .05$ level) and also to exceed a quarter of a standard deviation of the outcome measure. The first requirement merely obliged us to base each finding on a large enough number of cases that it would be unlikely to result from the differences among members of the groups being compared, rather than from differences among the groups; this criterion has nothing to do with high or low standards of accomplishment. Overall, 30% of the effects were "significant."

The second criterion, which we have called "materiality" or "educational significance," arose from a consensus early in the project that mean test score differences that amounted to less than a quarter of a standard deviation might well be "significant," thanks to large samples, but that they would not be educationally interesting. Figure 6 displays a somewhat idealized picture of a difference of this magnitude. We decided, in effect, that if the FT and NFT distributions coincided even more closely than Figure 6 shows, we were willing to ignore the differences among them. Overall, 49% of the effects were "material."

As further evidence of the reasonableness of this thoroughly arbitrary, judgmental standard, we note that a quarter of a standard deviation corresponds roughly to about two months' progress in terms of grade equivalents. Over a three-year or four-year period of intervention, a program really ought to generate a couple of months' relative progress; if not, it has not made much of a dent in the problem that the outcome measure is intended to document. A year would be too much to ask; a week, too little. In our judgment, two months is a comfortable criterion. The reader who wishes to impose a more or less stringent standard can do so by reference to the detailed results of our computations of effects, presented in Volumes IV-B and IV-C of the report, cited above.

4. Could attrition have led us to mistaken conclusions? Our analytic sample comprises only a minority of the children who were pretested and to whom the Follow Through "treatments" were applied. What of the others? Did

Figure 6 Two normal distributions with the same variance and means separated by 0.25 standard deviation, illustrating the minimum "educationally-significant" FT/NFT difference.



they leave because they were in some relevant sense different from those who stayed? Had they all stayed, how would they have performed on the outcome tests, and would our measure of program effectiveness have turned out differently? The pluperfect subjunctive is unfortunately and notoriously unresearchable. Our closest approach to answers to such questions derives from a series of studies of the attrition question which we performed when the extent of developing attrition become apparent. From the results of these studies, we conclude, to our relief, that the same kinds of children left the FT and NFT groups, site by site, at least with respect to income and pretest scores. We therefore feel reasonably confident that attrition has not done serious damage to our comparisons of FT with NFT. Our results apply less well, it is true, to the more mobile segment of the disadvantaged population than to those who tend to stay in one place. But then the Follow Through world was not sampled probabilistically from any larger population: it is simply an availability sample, clustered by more or less relevant political criteria, of variably disadvantaged children and the adults associated with them. All anybody can do to generalize from such a "sample" is to describe it thoroughly, as we have done in Chapter Three of Volume IV-A, and then to surmise cautiously that "this is the kind of thing that happens when these treatments are applied to these sorts of children." We claim neither more nor less generality for our findings.

5. <u>But what if some of the models were not adequately implemented?</u> Is it fair, after all, to judge the effectiveness of an educational strategy by its poor performance in a school system that bungled the model in practice, perverted it, or simply ignored it? Do we really know how powerful the models are until we have at least ensured that they have been implemented?

Perhaps our clearest finding indicates that local circumstances, attitudes and activities, measureless to this evaluation, have much more to do with the effectiveness of Follow Through models, without exception, than have the theoretical characteristics of the models. As this pattern began to be clear in the early years of the evaluation, we turned our primary attention from the study of model average effects to that of local site group effects, so as not to obscure the results of good implementation averaging them with those of bad ones. Most of the models were tried out in

enough sites that one might reasonably expect at least one good implementation: if a model did not make its point in any of its sites, however powerful it might be in principle, one might legitmately wonder about its practical value.

In a practical sense, we can be fairly confident that a model can produce at least as much progress as it showed in its best site. On the other side of the coin, the same model can clearly do at least as badly as it did in its worst site. Lacking information about implementation (what if a model did best in a site that had not implemented it well?), we can only give the models the benefit of the doubt in this matter. If future evaluations of externally-guided compensatory interventions wish to be able to attribute outcomes more confidently to the specific kinds of intervention applied, they will have to take steps to find out, systematically, what local people did with the models and other associated program resources. Model labels have turned out to carry only a small fraction of the relevant information.